SCIENCE AND TECHNOLOGY TEXT MINING: STRUCTURED PAPERS

By

Dr. Ronald N. Kostoff
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217
Phone: 703-696-4198
Fax: 703-696-4274
Internet: kostofr@onr.navy.mil

Dr. James Hartley
Department of Psychology
Keele University
Staffordshire
ST5 5BG
UK

(*THE VIEWS IN THIS REPORT ARE SOLELY THOSE OF THE AUTHORS, AND DO NOT NECESSARILY REPRESENT THE VIEWS OF THE DEPARTMENT OF THE NAVY, ANY OF ITS COMPONENTS, OR THE UNIVERSITY OF KEELE*)

KEYWORDS: Science; Technology; Technical Literature; Research; Database; Title; Keywords; Abstract; Full Text References; Bibliography; Text Mining; Bibliometrics; Scientometrics

INTRODUCTION

Structured technical papers are full text manuscripts that contain a threshold number of canonical fields. Structured database papers are the database representations of the structured technical papers. They also contain a threshold number of canonical fields, using standardized formats where practical, but do not contain the full text. This document presents the case for instituting both the structured technical paper and the structured database paper, and shows how these documents would help accelerate the progress of science and technology.

Science and technology are the cornerstones of modern economies and militaries. Both the building blocks, and the products, of science and technology are knowledge and understanding. These building blocks and products are expressed in many forms, including hardware, software, trained personnel, and multi-media outputs (video, audio, data, text). However, the central unifying factor in the research planning, execution, and transition cycle is the technical literature. This literature binds together the various disciplines of research across space and time, and forms the foundation of future research.

The technical literature consists of technical papers, books, memos, letters, and in the modern age, myriad documents in text form on electronic media. Central to the literature is the peer-reviewed technical paper. This document provides the prospective reader some assurance that a threshold level of quality has been achieved through the expert peer review process. From the author's perspective, this document archives for posterity the lessons and insights learned from the conduct of research. These thoughts are expressed in a format most comfortable to the author, in order to communicate the knowledge most smoothly to the intended audience. From the user's perspective, the paper should be the source of requisite information that allows optimal performance of the user's function. Such functions may include research oversight, sponsorship, management, performance, transition, technology and engineering development, technical and military intelligence, and the conduct of other commercial and military operations that have some dependency on the products of research.

Unfortunately, the information that the author of a technical paper chooses to provide may not be wholly compatible with the needs of the user. This problem is further compounded when the technical paper is represented by a summary of fields in a large database (e.g., Science Citation Index, Engineering Compendex, MEDLINE), where all the key elements of the technical paper may not have been extracted into the database. Such a database is many times the only knowledge source from which many users will extract the information necessary to perform their function. For maximal information transfer between research performer and user it is imperative that the technical paper be structured to contain the information of interest to the user, and that the database with which the user interfaces to obtain this information contains the requisite information fields in an easily accessible format.

The above scenario leads to a number of questions. Why should the author of a technical paper modify the paper's structure to satisfy the needs of a user, other than purely intrinsic motivation? How many different users' needs should be taken in account when considering the output of a paper, and how are competing users' needs resolved? What are the priorities among users needs, and how should they determine the contents of a technical paper?

We address this situation from the perspective of a sponsor – performer relationship. Research, especially fundamental research, constitutes the bulk of the high quality peer-reviewed technical literature. The major sponsors of fundamental research globally are the respective sovereign governments, and the significant sponsors of applied research and technology are the sovereign governments as well. The national governments are footing the bill for the research that makes the paper and journals and technical databases possible, and these governments should therefore be setting requirements for the information that they require from the papers, journals, and databases in order to perform their oversight functions.

Yet, this is not the case today. For the most part, the authors determine the structure and content of the papers, the journals impact the content through the peer review process very weakly, and the databases only extract selected fields to present to the user. In what other sponsor – performer relationship would the performer determine what product the sponsor receives?

Compounding the paper limitation problem described above is the fact that only a very modest fraction of performed research ever gets documented. The disincentives (proprietary research, classified research, research focused on uncovering or correcting product problems, time spent for documentation removes time available for research, transitions tend to be rewarded at the expense of documentation) for documentation tend to outweigh the incentives. In summary, from the sponsor's perspective, substantial funds are allocated to the conduct of research, only a fraction of the research performed gets documented, and the information contained in the documentation leaves much to be desired. This information deficiency retards the progress of research performance, transition, oversight and evaluation.

The present report addresses the second of the problems above, namely, how should the technical paper be structured to provide primarily the sponsor, and secondarily the other users, with the information required for him/her to perform his/her function properly. The first problem of insufficient research documentation has been addressed elsewhere, and is of such a large magnitude that it will require the cooperation of sovereign governments worldwide to correct.

APPROACH

The structured technical paper, and its abridged representation in the large databases, should contain sufficient information to achieve the following objectives. The structured technical paper should accurately reflect both the approach used to accomplish the research and a deep understanding of the findings. The structured database paper should reflect the main concepts in the structured technical paper. A person searching the structured database paper would be able to retrieve this paper if any of the paper's main concepts are aligned with the interests expressed in the search query.

The structured technical paper and structured database paper have two types of fields, which can be subsumed under the general headings of bibliometric and text. The bibliometric fields include author, address, institution, country, sponsor (sometimes), and journal. The text fields include title, abstract, keywords, full text, and references. It is our thesis that minimum levels of information should be contained in each of these categories in the structured technical paper, and transferred (with the exception of the full text) to the structured database paper as well. In the future, if databases evolve into repositories of full text papers, the minimum requirements below would still apply. The following sections describe these minimal information requirements. Emphasis will be placed on describing the text field requirements, but those for the bibliometrics fields will be summarized initially.

I. Bibliometric Fields

1. Author – The author fields in all papers and database representations should contain the names of all authors, as all papers and most databases do presently. Unfortunately, while there is uniformity of

surname spelling, the first name tends to be treated differently in different databases. Some databases provide the full first name, some provide a first initial only. Also, some databases/ journals/ papers provide a middle initial; some don't. These different author name representations complicate bibliometric statistics when different databases are combined, and the end result is that authors tend to be under-represented in bibliometric statistics. Equally serious, if a database is being searched by author name, the paper may be overlooked entirely because of the abbreviations. A common standard for the treatment of full names needs to adopted, and should be decided at a community – representative workshop for addressing technical paper structure issues.

More seriously is the issue of common names. For bibliometrics purposes, authors with common names (J. Smith, C. Lin, J. Kim) cannot be uniquely identified. In this case, authors tend to be over-represented in bibliometrics statistics. Equally serious, retrievals based on author names will retrieve excessive documents. To eliminate this confusion, each author should be assigned a unique number, such as a social security number, that would be retained ad infinitum.

2. Journal – In the technical paper, the journal name is usually listed prominently, and in full. In the larger database, the journal may be listed in full, or abbreviated. The abbreviations may vary from journal to journal. As in the case of the author field, when different databases are combined, bibliometrics become complicated and journals may be under–represented unless detailed hand corrections are made.

3. Address/ Institution – Address and Institution fields are difficult to separate. There is little uniformity from within the different journals in the same database, much less different databases. Address/ institution fields tend to differ in the number of levels used for descriptive purposes, making the determination of numbers of papers from a given address difficult to compute. For example, one database may list an author's address as Harvard University, another may add Department of Chemistry, another may add Institute of Nanoscience, and a fourth may add Atomic Force Microscopy Laboratory. The addresses will only be added for bibliometric

purposes at the highest level (at best), with substantial detail lost on the process.

4. Country – There should be a standard list of country names.

5. Sponsor – For government users especially, sponsor information is extremely important when doing productivity evaluations using bibliometrics. All technical papers and databases should contain sponsor information and uniform sponsor descriptions.

Text Fields

Different text fields are also important from an information retrieval perspective. Insufficient text in these fields makes the paper invisible to the search engines. Not only are the potential users penalized by lack of access to this information due to insufficient text in the text fields, but authors reduce the chances of their papers being accessed and therefore cited, and consequently journals will have their Impact Factors reduced.

1. Title – The title should be a concise summary of the paper's contents. Fancy titles should be avoided.

2. Abstract – The abstract is the most important text field in the database, and is addressed at length in Appendix 1.

3. Keywords – Keywords are an important text field in the database, and are addressed at length in Appendix 2.

4. Full Text – The full text of the technical paper should contain at least all the fields addressed in the Abstract. While the full text does not need to contain all the words in the Keywords fields, since some of the Keywords could be meta – level words placing the paper in a larger context, the full text should somehow relate to concepts represented by each of the Keywords used.

5. References – The references field is a hybrid between text and bibliometrics. Papers and journals are not uniform in their treatment of references, and the databases are not uniform in their incorporation of references. References (citations) tend to be under-

represented in bibliometrics studies because of this non-uniformity. All databases should include references in the same format.

In summary, the bibliometric fields need to be standardized in terms of content and format, and the text fields need to have minimum content required for canonical fields. The databases need to include canonical fields, and have uniform formats.

A workshop representing a broad cross–section of the relevant user, performer, journal, and database communities needs to be convened to set requirements for these field contents and formats. This workshop should recommend:

- A common standard for journal names in major databases: we recommend that journal titles are given in full and not abbreviated;
- Criteria for standardizing authors' and institutions' addresses, including post-codes;
- A standard list of country names;
- Uniform sponsor descriptions; and
- A common standard for setting the references among papers, journals and databases.

APPENDIX 1 – STRUCTURED ABSTRACTS

The widespread use of multi-discipline technical databases - such as the Science Citation Index (SCI), MEDLINE, INSPEC, and the Engineering Compendex - for the generic purpose of Technology Watch [1] - has expanded the potential for increased technology transfer and cross-discipline innovation. Each record in these databases contains a number of fields that provide different levels of detail about the underlying full-text article (e.g., Title, Abstract, Keywords). The critical path to information transfer lies in the quality of the most detailed record field - the Abstract.

Yet there is a substantial lack of uniformity in the presentation of the information contained within the Abstracts in the technical literature. The records of research and review articles that contain Abstracts vary substantially in the volume of information they present, the categories of information they address, and in the clarity of their presentation.

We have used both medical and technical literatures extensively in our work, coming from the different perspectives of text mining [RNK], and psychology [JH]. Collectively, we have examined many thousands of Abstracts in myriad technical disciplines. When reading technical journal Abstracts, we have not always been able to identify one or more of the following: 1) the context of the research; 2) the purpose behind many of the articles; 3) the research approach; 4) the results obtained; 5) the conclusions reached, and 6) the potential applications of the research described.

However, we do *not* find these problems in the bulk of the *medical* research literature. Many medical research journals require that their authors address canonical categories in a common sequence under a series of sub-headings in their *Structured* Abstracts. The purpose of having such Structured Abstracts is to insure that sufficient data are presented systematically to satisfy the information requirements of different journal readers. (Appendix 1A gives an example of an unstructured technical Abstract and its structured version. Appendix 1B gives a variety of representative structured medical abstracts, with length and category requirements based upon unique journal needs.)

What are these common information requirements among different reader groups? For both research and review papers, most readers are interested in:

1. Why is the research important? (Background)
2. What is the purpose of the research? (Objectives)
3. What techniques are used in the conduct of the research or the conduct of the review? (Approach)
4. What new information is provided by the research or review? (Results) and
5. What conclusions can be drawn from the research or review? (Conclusions).

Different reader groups may also have additional information requirements, depending on their study objectives. Some groups may require additional categories to the five mentioned above, and some may require additional amounts of explanation for any of the categories presented.

For example, readers interested in technology transfer may require a category describing potential applications, as seen by the article's author(s). To take another example, readers unfamiliar with the paper's discipline may require a more readable jargon-free description of each category's contents. And, as a final example, evaluators might not only be interested in all of the categories above, but also find comments on the innovation and significance of the research results to be highly useful.

About a decade ago, the medical research community began implementing Structured Abstracts to address their unique requirements. A foundational paper [2] recommended that Abstracts contain the following categories: 1) *Research papers* - Objective, Design, Setting, Patients, Interventions, Main Outcome Measures, Results, Conclusions; 2) *Review papers* - Objective, Data Sources, Study Selection, Data Extraction, Data Synthesis, Conclusions. Different variants of these categories were implemented in many of the different medical journals.

The experience of the medical community with Structured Abstracts has been well documented [3, 4, 5]. In summary, Structured Abstracts tend to be longer than unstructured ones but no negative impact on creativity or originality has been identified. Evaluators have found the information content of Structured Abstracts to be more useful than that in unstructured

ones, and Structured Abstracts are now widely accepted in the medical literature as a positive improvement.

Our own experience of reading Structured and Unstructured Abstracts has convinced us there is no comparison. For text mining, or discipline research and evaluation, Structured Abstracts have substantially greater value. In fact, the benefits are so obvious that we have trouble understanding why Structured Abstracts have not yet been implemented in technical journals.

We recommend that the editors of technical journals convene to establish formats and guidelines for Structured Abstracts. As a starting point, we offer the following recommendations for both original research and review articles.

All disciplines should require the generic categories of Background, Objectives, Approach, Results, and Conclusions. A category of Potential Applications would be optional. Each journal could establish sub-categories to accentuate information of value to its unique discipline. For example, the Journal of the American Medical Association has established the following sub-categories for 1) *Original research articles*: Context, Objective, Design, Setting, Patients (or Participants), Interventions (include only if there are any), Main Outcome Measure(s), Results, and Conclusions; and 2) *Review articles*: Context, Objective, Data Sources, Study Selection, Data Extraction, Data Synthesis, and Conclusions. These sub-categories fit within the generic recommended categories, and contain specific requirements unique to patient studies. Specific examples of Structured Abstract guidance to authors can be found in [6], [7], and [8].

In addition, all text fields in the record should contain the same type of information, albeit at a different level of resolution. The need for this requirement was shown clearly in a recently published text mining study of Aircraft science and technology (S&T) [9]. Computational linguistics analyses of the Abstract and of the Keyword fields from a large number of aircraft-related records of the SCI showed that a very different perspective of Aircraft S&T was obtained from each field's analysis. This has far-reaching applications for information retrieval and discipline evaluation.

The most common criticism raised by editors and editorial committees concerning the suggestion that they implement Structured Abstracts is

based on space/cost grounds. It is true that Structured Abstracts are usually longer than unstructured ones. Nonetheless, most journals start their new articles on a fresh (right-hand) page – so the space is available – and this issue does not arise, of course, with electronic journals. Further, the issue of *cost-effectiveness* needs to be addressed. More informative Abstracts are likely to encourage greater readership, greater citation rates, better author bibliometric statistics, and higher journal Impact Factors.

In conclusion, we believe that the time has come for technical journals to implement the use of Structured Abstracts. The benefits to technology transfer, cross-disciplinary innovation, research review and evaluation, and to corporate and national security intelligence are likely to be substantial.

REFERENCES

1. Kostoff, R. N. (2003) "Text mining for global technology watch", Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799.

2. Ad Hoc Working Group for Critical Appraisal of Medical Literature. (1987) "A proposal for more informative abstracts of clinical articles". Ann. Intern. Med. 1987; 106: 598-604.

3. Taddio, A., Pain, T., Fassos, F. F., et al, (1994) "Quality of nonstructured and structured abstracts of original research articles in the British Medical Journal, the Canadian Medical Association Journal and the Journal of the American Medical Association". Canad. Med. Assoc. J. 1994; 150:1611-1615.

4. Hartley, J. (2000) "Clarifying the abstracts of systematic reviews". Bull. Med. Library Assoc. 2000; 88: 332-337.

5. Hartley, J. (2004) "Current research on structured abstracts". Paper submitted for publication. Copies available from the author.

6. Haynes, R.B., Mulrow, C. D., Huth, E. J., et al, (1990) "More informative abstracts revisited". Ann. Intern. Med. 1990; 113: 69-76

7. International Committee of Medical Journal Editors (1997) "Uniform requirements for manuscripts submitted to biomedical journals". 1997; 336: 309-315.

8. Notes to contributors. Brit. J. Educ. Psychol. 2001; 71 (Inside back cover page).

9. Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. (2000) "Database tomography applied to an aircraft science and technology investment strategy". Journal of Aircraft, 37:4, July-August 2000. Also, see Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. A., "Database tomography applied to an aircraft science and technology investment strategy", TR NAWCAD PAX/RTR-2000/84, Naval Air Warfare Center, Aircraft Division, Patuxent River, MD.

APPENDIX 1A - An example of a Traditional Abstract and its Structured version.

To achieve the structured version, the text was re-arranged to fit the standard headings, and additional detail was provided. (Abstract reproduced with permission of the author.)

1) Traditional version

This paper describes two novel complementary approaches for systematically enhancing the process of innovation and discovery. One approach is workshop-based and the other approach is literature-based. Both approaches have the common features of exploring knowledge from very disparate technical disciplines and technologies, and transferring insights and understanding from one or more disparate technical areas to another. It is highly recommended that the approaches be combined into a single process. The integrated approach has the potential to be a major breakthrough for the systematic promotion of innovation and discovery.

2) Structured version

Background: One important factor in innovation is the transfer of information and understanding developed in one or more disciplines to other, sometimes very disparate, disciplines.
Objectives: The aim of this research was to develop and demonstrate a systematic method for enhancing cross-discipline knowledge transfer that overcomes the limitations of existing literature and workshop-based approaches.
Approach: Both the traditional literature and workshop-based approaches were re-conceptualized and combined in a two-stage three-phase procedure – a literature-based discovery stage followed by a three-phase workshop stage (a two-month e-mail-facilitated pre-meeting phase, a two-day workshop phase, and a post workshop e-mail phase) – on the topic of Autonomous Flying Systems.
Results: The revised literature and pre-meeting approach was an excellent vehicle for identifying (i) a broad range of disciplines that supported the central theme and were represented at the workshop, (ii) promising concepts to pursue, and (iii) leading experts in the diverse disciplines to participate in the workshop. Many ideas were developed further at the

workshop, and one outcome was a proposal concerning future research opportunities for Autonomous Flying Systems.

Conclusions: Tandem literature and workshop stages are required if maximum innovation stimulation is to be obtained. Substantial planning is required if the right combination of disciplines is to be represented at the workshop. Active facilitation of discussion during the two e-mail phases is crucial to provide concept enhancement.

APPENDIX 1B - Samples of Structured Medical Abstracts.

Example 1

Background: The cause of pain in osteoarthritis is unknown. Bone has pain fibers, and marrow lesions, which are thought to represent edema, have been noted in osteoarthritis.
Objective: To determine whether bone marrow lesions on magnetic resonance imaging (MRI) are associated with pain in knee osteoarthritis.
Design: Cross-sectional observational study.
Setting: Veterans Affairs Medical Center.
Patients: 401 persons (mean age, 66.8 years) with knee osteoarthritis on radiography who were drawn from clinics in the Veterans Administration health care system and from the community. Of these persons, 351 had knee pain and 50 had no knee pain.
Measurements: Knee radiography and MRI of one knee were performed in all participants. Those with knee pain quantified the severity of their pain. On MRI, coronal T-2-weighted fat-saturated images were used to score the size of bone marrow lesions, and each knee was characterized as having any lesion or any large lesion, The prevalence of lesions acid large lesions in persons with and without knee pain was compared; in participants with knee pain, the presence of lesions was correlated with severity of pain.
Results: Bone marrow lesions were found in 272 of 351 (77.5%) persons with painful knees compared with 15 of 50 (30%) persons with no knee pain ($P < 0.001$). Large lesions were present almost exclusively in persons with knee pain (35.9% vs. 2%; $P < 0.001$). After adjustment for severity of radiographic disease, effusion, age, and sex, lesions and large lesions remained associated with the occurrence of knee pain. Among persons with knee pain, bone marrow lesions were not associated with pain severity.
Conclusions: Bone marrow lesions on MRI are strongly associated with the presence of pain in knee osteoarthritis.

Example 2

Purpose: Congestive heart failure is an important cause of patient morbidity and mortality. Although several randomized clinical trials have compared beta -blockers with placebo for treatment of congestive heart failure, a meta-analysis quantifying the effect on mortality and morbidity has not been performed recently.
Data Sources: The MEDLINE, Cochrane, and Web of Science electronic databases were searched from 1966 to July 2000. References were also identified from bibliographies of pertinent articles.
Study Selection: All randomized clinical trials of beta -blockers versus placebo in chronic stable congestive heart failure were included.
Data Extraction: A specified protocol was followed to extract data on patient characteristics, beta -blocker used, overall mortality, hospitalizations for congestive heart failure, and study quality.
Data Synthesis: A hierarchical random-effects model was used to synthesize the results, A total of 22 trials involving 10 135 patients were identified. There were 624 deaths among 4862 patients randomly assigned to placebo and 444 deaths among 5273 patients assigned to beta -blocker therapy. In these groups, 754 and 540 patients, respectively, required hospitalization for congestive heart failure, The probability that beta -blocker therapy reduced total mortality and hospitalizations for congestive heart failure was almost 100%. The best estimates of these advantages are 3.8 lives saved and 4 fewer hospitalizations per 100 patients treated in the first year after therapy. The probability that these benefits are clinically significant (>2 lives saved or >2 fewer hospitalizations per 100 patients treated) is 99%, Both selective and nonselective agents produced these salutary effects. The results are robust to any reasonable publication bias,
Conclusions: Beta -Blocker therapy is associated with clinically meaningful reductions in mortality and morbidity in patients with stable congestive heart failure and should be routinely offered to all patients similar to those included in trials.

Example 3

Purpose. Our aim was to compare the role of remote afterloaded high-dose-rate brachytherapy (HDRB) with traditional low-dose-rate brachytherapy (LDRH) for patients with invasive primary vaginal carcinoma.
Methods. The study group comprised 190 patients with invasive carcinoma of the vagina. The patients were staged according to the International Federation of Gynecology and Obstetrics (FIGO) staging system. Eighty patients were treated with intracavitary high-dose rate iridium 192 brachytherapy with or without external beam therapy These patients are compared with a historical group of 110 patients treated with intracavitary low-dose-rate radium 226 or cesium 137 brachytherapy with or without external beam therapy
Results. No significant differences were found for stages, tumor grade or location between the two groups. Crude 5-year survival for all patients was 41% in the former LDRB group, 81% in stage I and 43% in stage II. Overall actuarial 3-year survival and disease-specific survival rates for all patients in the HDRB series were 51% and 66%, respectively. Disease-specific 3-year survival attained 83% in stage I and 66% in stage II. There were no significant differences in local and distant recurrences between the treatment modalities. The comparison of treatments with or without external beam radiation and of complications showed no significant differences between the HDRB and LDRB series.
Conclusion. With HDRB and its advantages of decreased radiation exposure and patient immobilization and precise positioning, treatment results to be obtained are at least similar to traditional LDRB for primary vaginal **cancer**.

APPENDIX 2 – STRUCTURED KEYWORDS

In this appendix, we consider the generation, use and value of key words in research articles designed to help readers, writers, indexers and abstractors locate related information.

1. Introduction

All research articles begin with a title. Most include an abstract. Many include 'key words', as illustrated at the beginning of this report. All three of these features describe an article's contents with varying degrees of detail and abstraction. The title is designed to stimulate the reader's interest. The Abstract summarizes the content. The half-dozen or so key words, sometimes called 'descriptors' or 'subject headings', indicate the main concepts and fields of concern (While 'key words' is the common usage, strictly speaking these descriptors should be called 'key phrases', since multi-word phrases can be used as descriptors in most publications). Today, all three elements (together with the name/s of the author/s) are required in any serious bibliographic database designed to aid electronic information retrieval.

Key words typically:

1. allow readers to decide whether or not an article contains material relevant to their interests;
2. provide readers with suitable terms to use in web-based searches to find other materials on the same or similar topics;
3. help indexers/editors group together related materials in, say, the end-of-year issues of a particular journal or a set of conference proceedings;
4. allow editors/researchers to document changes in a subject discipline over time (although not everyone agrees with this: see [1]); and
5. link the specific issues of concern to issues at higher meta (abstraction) levels (See Appendix 2A for examples of meta-level key word usage).

2. Who uses key words?

Key words have appeared in technical journal articles for decades. Yet, the research literature in the field provides no clear indication of who uses key words and why. Table 1 shows the percentage of journals using key words in different areas and disciplines. It can be seen that there are disciplinary

differences (there is less use of key words in English than, say, Psychology). But even within disciplines there is much heterogeneity. There appear to be no formal requirements for key words, no rules for formulating them, little guidance on how to write them (but see Appendix 2B), and no instructions for reviewers on how to assess them. This is surprising in view of the fact that, presumably, a wise choice of key words increases the probability that a paper will be retrieved and read, thereby potentially improving author citation bibliometrics and journal Impact Factor.

Table 1

The approximate percentages of research journals in different areas and disciplines supplying key words

Arts	Education	Psychology	Science	Medicine	Statistics
5	20	30	50	50	75

3. What are the advantages of key words?

While writing this report, we wrote to 35 editors of journals that use key words. We asked them about their practice relative to key words, and about what they perceived to be advantages and disadvantages of key words. Table 2 summarizes the main responses of the 22 editors who replied. This table shows that there is considerable diversity of opinion but that, generally speaking, these editors perceive more advantages for key words than disadvantages.

Table 2

The advantages and disadvantages of supplying key words as perceived by 22 editors

Advantages

- They make it easier for people to do electronic literature searches.
- Readers can use them to look up relevant articles in the index.
- Readers and researchers can quickly and easily locate particular articles within their area of interest.
- They are useful for abstracting and indexing services.

- They help editors glean from the authors those things about their papers that they consider critical in terms of relating it to the broader literature in the discipline.
- They can help editors prepare the Index at the end of each volume.
- They can provide editors with a way of tracking the coverage of articles in the journal, both currently and over time.
- They can be useful for assigning papers to reviewers.

Disadvantages

- Relevant articles may be missed if the author doesn't use the right key words. Currently authors do not appear to give much thought to their importance for information retrieval.
- If the key words are not accurate or general enough, they can mislead as well as help.
- Authors sometimes use key words to make wider claims for their papers than the content justifies.
- Key words sometimes seem inappropriate (e.g. in Arts journals) where the field is so diverse, where authors may be talking about their work in a personal way, and/or where they may be using arcane distinctions.
- Authors sometimes forget to include them, and this causes more work.

4. Who chooses the key words?

Table 3 indicates that there are several different ways in which key words are chosen. The most common method (over 50%) is for the authors to supply as many words as they choose (within bounds), but sometimes a specified number of words is required (often about six). The next main method (about 20%) is for the authors to choose key words that fit into categories already prescribed by the journal's 'instructions to authors'. Thus, for example, whoever generates the key words for medical articles must select only words from the MeSH (Medical Subject Headings) Taxonomy – a structured taxonomy used by MEDLINE. Similarly, lists of appropriate words are provided by *Biological Abstracts* and *Chemical Abstracts*. In these situations the number of words allowed, and the numbers of categories from which to choose, can vary. Some psychology journals, for example, allow authors to list key words from any of the 5,000 terms that appear in the American Psychological Society's *Thesaurus of Psychological Index Terms.* Finally some key words are generated automatically at proof stage (as in the *Journal of Information Science*).

Table 3

Different methods for supplying key words.

- Authors supply them with no restrictions on the numbers allowed.
- Authors supply up to a fixed number (e.g., six).
- Authors supply key words as appropriate from a specified list.
- Editors supplement/amend authors' key words.
- Editors supply key words.
- Editors supply key words from a specified list.
- Referees supply key words from a specified list.
- Key words are allocated according to the 'house-rules' applied to all journals distributed by a specific publisher.
- Key words are determined by computer program (e.g., LISA) at proof stage.

The editors surveyed had different views about the best procedures for creating key words, and these seemed to relate to the methods used by their journals. One who supported authors generating their own key words said, 'Who but the author could possibly do a better job of picking the key words for an article?' Another, however, had found that the list of author-generated key words for his journal was 'growing out of hand', so he created a 'standard list' for future authors to use as far as they found it possible. Similarly, a third editor remarked, 'In order to be systematic it is much better to have a 'closed' set of key words rather than letting everyone come up with their own.' These differences about the best sources of key words parallel similar ones about the generation of book indexes [see http://www.asindexing.org].

Actually, there is research showing that, although key words are typically chosen by authors, such people are not very good at generating descriptors of their research [2]. Furthermore, different authors can disagree in their choice of key words for a particular paper [3]. Even worse, trained indexers are not much more reliable [4,5,6]. So, limiting the words allowed for particular articles may be useful - in the sense that there will be less variability – but some concepts may get omitted. We know of no research comparing author-generated with machine-generated key words in this respect but we note, from our experience of using Natural Language Processors to generate multi-word phrases from technical articles, that

many key technical phrases are missed and many non-technical phrases are included in these circumstances.

The reader may be interested to compare they key words supplied by the authors for an article to be published in the *Journal of Information Science* [7] with those generated automatically at proof stage. The words supplied by the authors were: descriptors, information retrieval, author-generated key words, key word searches, disciplinary differences, and indexing. The words supplied by the Journal were: keywords, periodicals, articles, information retrieval, relevance, and science and technology. These differences suggest that it might be useful to use both systems – author and automated - and to then decide on the final selection.

5. Can key words be made more useful?

As noted above, key words are valuable for authors, readers, indexers, abstractors, and people who search for related information on the web and in research databases. Undoubtedly, there would be greater consistency between authors and journals within different disciplines if standard terms were used within various categories. And there would probably be more uniform information retrieval across different disciplines if searchers used standardized terms and categories.

In addition, independently of how key words are chosen (as described above) a comprehensive listing of all the key words in a database should be made available periodically for the database's users. With MEDLINE, for example, searchers can browse the MeSH taxonomy before developing a query. If a term is in MeSH, the user will know that some MEDLINE records will be accessed with use of that term in the search engine, although the number of 'hits' will not be known with the present Mesh structure. It would be additionally helpful if the number of records containing a particular key word could also be displayed. Such a procedure would allow the analyst to know the number of hits beforehand, and would take a lot of mystery out of the searching process.

One of the editors surveyed suggested that, with future developments, all of the issues discussed above are really non-problems. As he put it, 'Inverted-full-text-Boolean indexing and online searching (with similarity algorithms and citation-ranking) will soon make keywords and human-subject-classification things of the past.' Put more simply, eventually we

will be able to input any words, pairs of words, or phrases that we like from an article into a search engine and come up with related materials.

This may be true, but it ignores the fact that using such an open-ended route might swamp one with information, and ignores the complexities involved in searching when the input query does not match the exact terminology used in the textual material to be retrieved.

If the query and the text being searched contain concepts at the same level of abstraction but with different specific terminology, then some sort of automated thesaurus or lexicon is required to translate between the two terminologies. For very mature disciplines, like the English language, such a thesaurus is doable. For rapidly changing disciplines characteristic of cutting-edge science, such a thesaurus would tend to lag the state-of-the-art unless a large staff were employed to maintain constant literature awareness and update the thesaurus immediately. In addition, many words/ phrases have multiple meanings (e.g., bank-repository of money, act of depositing money, depend, side of river, airplane maneuver), and some auxiliary agent would be required to identify the thesaurus term appropriate for the context.

If the query and the text being searched contain concepts at different levels of abstraction, then some sort of translator across meta-levels would be required. For this process, using humans appears to be a necessity, at least in the forseeable future.

Indeed, looking to the future, what we think is required are more specific criteria for determining key words. Authors might thus be given a list of categories or subheadings (as in structured abstracts) indicating different levels of text and asked to supply key words for each one (as appropriate). Table 4 provides a specific example for a technical article. General guidelines for selecting key words, such as those in Appendix 2, should also be supplied. For less technical articles, only a sub-set of the categories in Table 4 would be required.

Table 4

An example of sub-headings or categories for authors to use in listing key words for a technical paper on measuring protein masses

Sub-headings	*Key words*
Directly related technologies	electrospray ionization, ion trap, chromotography magnetic sector
Indirectly related technologies	biochemistry, Fourier transforms, statistical averaging, chemistry
Technical phenomena	collisions, dissociation, multiply-charged states, affinity, ionization
Capabilities addressed	determine structures, study disease evolution, growth rates
Potential applications	drug screening, pathology, cloning, clinical testing
Instruments and Procedures	liquid chromatograph, quadrupole selector, tandem mass spectrometer, molecular beam, skimmer
Theoretical tools	time series analysis, ab initio electron orbitals solver, cluster analysis package

Note that while some of this information might be contained in the full text title, abstract and text, some of it may not be.

Acknowledgements. We are grateful to Edward Gbur and Bruce Trumbo for helpful comments on the appendix.

References

1. Leydesdorff, L. (1997) Why words and co-words cannot map the development of the sciences, *Journal of the American Society for Information Science,* 48 (5) 418-427.

2. Foltz, P. W. Improving Human-Proceedings interaction: Indexing the CHI Index. (Paper available from the author, pfoltz+@pitt.edu)

3. Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987) The vocabulary problem in human-system communication, *Communications of the ACM,* 30 (11) 964-971.

4. Tarr, D. and Borko, H. (1974) Factors influencing inter-indexer consistency, *Proceedings of the ASIS 37th Annual Meeting,* 11 50-55 [cited in 2].

5. Jones, K. P. (1983) How do we index: A report of some Aslib Informatics Group activity, *Journal of Documentation,* 39 (1) 1-23.

6. Kobayashi, M. and Takeda, T. (2000) Information retrieval on the Web, IBM Research report, RT0347, April. http://citeseer.nj.nec.com/kobayashi00information.html

7. Hartley, J. and Kostoff, R. N. (2003) How useful are 'key words' in scientific journals, *Journal of Information Science,* 29 (5) (in press).

8. Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. (2000) Database Tomography Applied to an Aircraft Science and Technology Investment Strategy. *Journal of Aircraft,* 37 (4) 727-730.

APPENDIX 2A – An Example of Differences in Key Words and Meta-Level Abstraction

In a text mining study of Science Citation Index journal articles on Aircraft Science and Technology [8], the Key Word field and the Abstracts field were mined independently. In the Key Word field, a group of high frequency key words was concentrated on longevity and maintenance; this view of the Aircraft literature was not evident from the high frequency phrases from the Abstract field, where lower frequency phrases had to be examined to identify thrusts in this mature technology area.

The contents of the Keyword field reflect summary judgements of the main focus of the paper's contents by the author or indexer, and represent a higher level description of the contents than the actual words in the paper or abstract. Thus, one explanation for the difference between the conclusions from the high frequency Keywords and Abstract phrases is that the body of non-maintenance Abstract phrases, when considered in aggregate from a gestalt viewpoint, is perceived by the author or indexer as oriented towards maintenance or longevity. For example, the presence of the material category phrase CORROSION in the Abstract could be viewed by the indexer as indicative of a maintenance-focused paper, since many maintenance problems are due to the presence of corrosion. Another explanation is that maintenance and longevity issues are receiving increased attention, and the authors/ indexers may be applying (consciously or subconsciously) this 'spin' to attract more reader interest.

As another example, the Abstract phrases from the Aircraft study contained heavy emphasis on laboratory and flight test phenomena, whereas there was a noticeable absence of any test facilities and testing phenomena in the Keyword field. Again, the indexers may view much of the testing as a means to an end, rather than the end itself, and their Keywords reflect the ultimate objectives or applications rather than detailed approaches for reaching these objectives. However, there was also emphasis on high performance in the Abstract phrases, a category conspicuously absent from the Keywords. The presence of descriptors from the mature technology or longevity categories in the Keywords, coupled with the absence of descriptors from the high performance category, provided a very different picture of the Aircraft research literature than did the presence of high performance descriptors and the lack of longevity and

maintenance descriptors in the Abstract phrases. Extrapolating these results to information retrieval, query terms applied to the Key Words field could retrieve different literatures from those same terms applied to the Abstracts field, at least for some technical disciplines.

APPENDIX 2B - Ten Ways to Select Effective Key Words and Phrases

Gbur and Trumbo (1995) published a list of ways of producing effective key words and phrases. It is repeated here (lightly edited) with their permission and that of *The American Statistician.*

1. Use simple, specific noun clauses. For example, use *variance estimation*, not *estimate of variance.*

2. Avoid terms that are too common. Otherwise the number of 'hits' will be too large to manage.

3. Do not repeat key words from the title. These will be picked up anyway.

4. Avoid unnecessary prepositions, especially *in* and *of.* For example, use *data quality* rather than *quality of data.*

5. Avoid acronyms. Acronyms can fall out of favour, and be puzzling to beginners and/or overseas readers.

6. Spell out Greek letters and avoid mathematical symbols. These are impractical for computer-based searches.

7. Include only the names of people as key words if they are part of an established terminology: e.g., *Skinner box*, *Poisson distribution.*

8. Include where applicable mathematical or computer-techniques, such as *generating function*, used to derive results, and a statistical philosophy or approach such as *maximum likelihood*, or *Bayes' theory.*

9. Include alternative or inclusive terminology. If a concept is, or has been known by different terminologies, use a key word that might help a user conducting a search across a time-span, or from outside your speciality. For example, the statistician's *characteristic function* is the mathematician's *Fourier transform.* And in some countries *educational administration* is *educational management.*

10. Note areas of applications where appropriate.

Source:
From Section 5: Suggestions for Authors: Selecting Key Words and Phrases, pp. 31-32 from "Key Words and Phrases - the Key to Scholarly Visibility and Efficiency in an Information Explosion" by E. E. Gbur & B. Trumbo, pp 29-33, Vol. 49, No.1, June 1995. Reprinted with permission of *The American Statistician.*